CMCL 2021 Shared Task on Eye-Tracking Prediction

Nora Hollenstein
University of Copenhagen
nora.hollenstein@gmail.com

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Cassandra Jacobs
University of Wisconsin
jacobs.cassandra.l@gmail.com

Yohei Oseki
University of Tokyo
oseki@g.ecc.u-tokyo.ac.jp

Laurent Prévot
Aix-Marseille University
laurent.prevot@univ-amu.fr

Enrico Santus
Bayer Pharmaceuticals
esantus@gmail.com

Abstract

Eye-tracking data from reading represent an important resource for both linguistics and natural language processing. The ability to accurately model gaze features is crucial to advance our understanding of language processing. This paper describes the Shared Task on Eye-Tracking Data Prediction, jointly organized with the eleventh edition of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2021). The goal of the task is to predict 5 different token-level eye-tracking metrics from the Zurich Cognitive Language Processing Corpus (ZuCo). Eye-tracking data were recorded during natural reading of English sentences. In total, we received submissions from 13 registered teams, whose systems include boosting algorithms with handcrafted features, neural models leveraging transformer language models, or hybrid approaches. The winning system used a range of linguistic and psychometric features in a gradient boosting framework.

1 Introduction/Overview

The ability of accurately modeling eye-tracking features is crucial to advance the understanding of language processing. Eye-tracking provides millisecond-accurate records on where humans look, shedding lights on where they pay attention during their reading and comprehension phase (see the example in Figure 1). The benefits of utilizing eye movement data have been noticed in various domains, including natural language processing and computer vision. Not only can it reveal the workings of the underlying cognitive processes of language understanding, but the performance of computational models can also be improved if their inductive bias is adjusted using human cognitive signals such as eye-tracking, fMRI, or EEG data (Barrett et al., 2016; Hollenstein et al., 2019; Toneva and Wehbe, 2019). Thanks to the recent introduction of a standardized dataset (Hollenstein et al., 2018, 2020), it is now possible to compare the capabilities of machine learning approaches to model and analyze human patterns of reading.

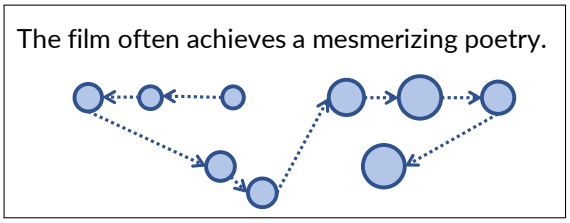


Figure 1: Example sentence from the ZuCo corpus read by a single reader. The blue dots mark fixations on the corresponding words above, a wider diameter represent a longer fixation duration.

In this shared task, we present the challenge of predicting eye word-level tracking-based metrics recorded during English sentence processing. We encouraged submissions concerning both cognitive modeling and linguistically motivated approaches (e.g., language models). All data files are available on the competition website.[1]

2 Related Work

Research on naturalistic reading has shown that fixation patterns are influenced by the predictability of words in their sentence context (Ehrlich and Rayner, 1981). In natural language processing and psycholinguistics, the most influential account of the phenomenon is surprisal theory (Hale, 2001; Levy, 2008), which claims that the processing difficulty of a word is proportional to its *surprisal*, i.e., the negative logarithm of the probabil-

[1] https://competitions.codalab.org/competitions/28176

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 72–78

ity of the word given the context. Surprisal theory was the reference framework for several studies on language models and eye-tracking data prediction (Demberg and Keller, 2008; Frank and Bod, 2011; Fossum and Levy, 2012). These studies use the data from the Dundee Corpus (Kennedy et al., 2003), which consists of sentences from British newspapers with eye-tracking measurements from 10 participants, as one of the earliest and most popular benchmarks.

Later work on the topic found that the perplexity of a language model is the primary factor determining the fit to human reading times (Goodkind and Bicknell, 2018), a result that was confirmed also by the recent investigations involving neural language models such as GRU networks (Aurnhammer and Frank, 2019) and Transformers (Merkx and Frank, 2020; Wilcox et al., 2020; Hao et al., 2020). Using an alternative approach, Bautista and Naval (2020) obtained good results for the prediction of eye movements with autoencoders.

In addition to the ZuCo corpus used for this shared task (see Section 4), there are several other resources of eye-tracking data for English. The Ghent Eye-Tracking Corpus (GECO; Cop et al., 2017) is composed of the entire Agatha Christie's novel *The Mysterious Affair at Styles*, for a total of $54,364$ tokens, it contains eye-tracking data from 33 subjects, both English native speakers (14) and bilingual speakers of Dutch and English (19), and comes with the Dutch counterpart. The Provo corpus (Luke and Christianson, 2017) contains 55 short English texts about various topics, with 2.5 sentences and 50 words on average, for a total of $2,689$ tokens, and eye-tracking measures collected from 85 subjects. Annotated eye-tracking corpora are also available for other languages, including German (Kliegl et al., 2006), Hindi (Husain et al., 2015), Japanese (Asahara et al., 2016) and Russian (Laurinavichyute et al., 2019), among others.

3 Task Description

In this shared task, we present the challenge of predicting eye-tracking-based metrics recorded during English sentence processing. The task is formulated as a regression task to predict the following 5 eye-tracking features for each token in the context of a full sentence:

1. nFix (number of fixations): total number of fixations on the current word.
2. FFD (first fixation duration): the duration of the first fixation on the prevailing word.
3. TRT (total reading time): the sum of all fixation durations on the current word, including regressions.
4. GPT (go-past time): the sum of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word.
5. FIXPROP (fixation proportion): the proportion of participants that fixated the current word (as a proxy for how likely a word is to be fixated).

The goal of the task is to train a model which predicts these five eye-tracking features for each token in a given sentence.

Feature	min	max	mean (std)
nFix	0.0	7.25	1.1 (0.7)
FFD	0.0	296.8	77.3 (34.4)
GPT	0.0	2424.9	154.1 (143.6)
TRT	0.0	996.2	128.8 (88.6)
FIXPROP	0.0	1.0	0.67 (0.26)

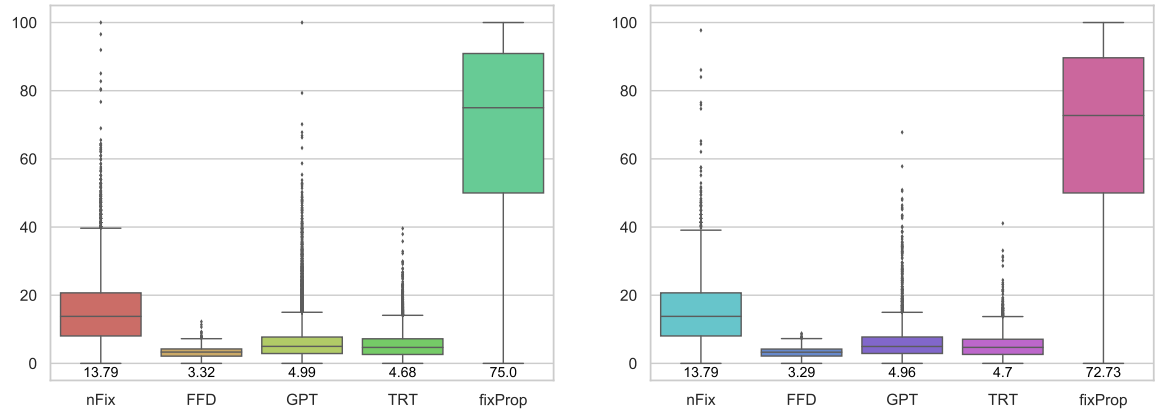
Table 1: Minimum, maximum, mean and standard deviation of the feature values *before scaling* in both training and test data, after averaging across readers.

Feature	min	max	mean (std)
nFix	0.0	100.0	15.1 (9.5)
FFD	0.0	12.2	3.2 (1.4)
GPT	0.0	100.0	6.4 (5.9)
TRT	0.0	41.1	5.3 (3.7)
FIXPROP	0.0	100.0	67.1 (26.0)

Table 2: Minimum, maximum, mean and standard deviation of the *scaled* feature values in both training and test data, after averaging across readers.

4 Data

We use the eye-tracking data recorded during normal reading from the freely available Zurich Cognitive Language Processing Corpus (ZuCo; Hollenstein et al., 2018, 2020). ZuCo is a combined eye-tracking and EEG brain activity dataset, which provides anonymized records in compliance with an ethical board approval and as such it does not



(a) Training data.

(b) Test data.

Figure 2: Boxplot showing the feature value distributions of both training and test sets. Below each box is the median value of each feature.

contain any information that can be linked to the participants.

The eye-tracking data was recorded with an EyeLink 1000 system in a series of naturalistic reading experiments. Full sentences were presented at the same position on the screen one at a time. The participants read each sentence at their own reading speed. The reading material included sentences from movie reviews from the Stanford Sentiment Treebank (Socher et al., 2013) and a Wikipedia dataset (Culotta et al., 2006). For a detailed description of the data acquisition, please refer to the original publications. An example sentence is presented in Figure 1.

We use the normal reading paradigms from ZuCo, i.e, Task 1 and Task 2 from ZuCo 1.0, and all tasks from ZuCo 2.0. We extracted the eye-tracking data from all 12 subjects from ZuCo 1.0 and all 18 subjects from ZuCo 2.0. The dataset contains 990 sentences. All sentences were shuffled randomly before splitting into training and test sets. The training data contains 800 sentences, and the test data 190 sentences.

4.1 Preprocessing

Tokenization The tokens in the sentences are split in the same manner as they were presented to the participants during the reading experiments. Hence, this does not necessarily follow a linguistically correct tokenization. For example, the sequences "(except," and "don't" were presented as such to the reader and not split into "(", "except", "," and "do", "n't" as a tokenizer would do. Sentence endings are marked with an <*EOS*> symbol added to the last token.

Feature Extraction The eye-tracking feature values are scaled between 0 and 100 to facilitate evaluation via the mean absolute error. The features NFIX and FIXPROP are scaled separately, while FFD, GPT and TRT are scaled together since these are all dependent and measured in milliseconds. The features are averaged across all readers. The data was scaled and randomly shuffled before splitting into training and test data. Tables 1 and 2 show the ranges of the eye-tracking features before and after scaling. Figure 2 depicts the feature value distributions in both training and test sets, showing that the distributions are very similar in both splits.

5 Evaluation

In this section, we describe the evaluation procedure used to assess the submitted predictions of the participating teams.

Any additional data source was allowed to train the models, as long as it is freely available to the research community. For example, additional eye-tracking corpora, additional features such as brain activity signals, pre-trained language models, etc.

5.1 Scoring Metric

The submitted predictions are evaluated against the real eye-tracking feature values using the mean absolute error (MAE) metric, a measure of errors between paired observations including comparisons of predicted (y) versus observed (x) values for each

Rank	Team Name	MAE	NFIX	FFD	GPT	TRT	FIXPROP	Reference
1	LAST	3.813	3.879	0.655	2.197	1.524	10.812	Bestgen (2021)
2	TALEP	3.833	3.761	0.662	2.180	1.486	11.076	Dary et al. (2021)
3	TorontoCL	3.929	3.944	0.671	2.227	1.516	11.286	Li and Rudzicz (2021)
4	LangResearchLab_NC	3.949	4.039	0.674	2.248	1.568	11.216	Agarwal and Chatterjee (2021)
5	CogNLP-Sheffield	3.957	3.956	0.689	2.260	1.529	11.349	Vickers et al. (2021)
6	OSU	3.977	3.987	0.682	2.364	1.540	11.311	Oh (2021)
7	MTL782_IITD	4.064	4.115	0.719	2.264	1.622	11.599	Choudhary et al. (2021)
8	KonTra	4.216	4.263	0.698	2.756	1.682	11.683	Yu et al. (2021)
9	Sabhay_Jain	4.257	4.264	0.848	2.476	1.721	11.974	-
10	ReadMe	4.383	4.363	0.741	2.502	1.761	12.549	Balkoca et al. (2021)
11	PIHKers	4.388	4.335	0.715	3.059	1.713	12.118	Salicchi and Lenci (2021)
12	ChiSquareX	4.676	4.557	1.281	2.810	2.289	12.445	-
-	MEAN BASELINE	7.357	7.303	1.149	3.782	2.778	21.775	-
13	IIIT_DWD	9.762	8.845	1.589	4.633	3.296	30.446	-

Table 3: Overall results showing the best submission per team and the mean baseline. The teams are ranked by the MAE averaged across all five eye-tracking features (third column).

word in the test set:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (1)$$

The winning system is defined as the one with the lowest average MAE across all 5 eye-tracking features.

5.2 Mean Baseline

We use the mean central tendency as a baseline for this regression problem, i.e., we calculate the mean value for each feature from the training data and use it as a prediction for all words in the test data. Table 3 shows the MAE scores achieved by this mean baseline for each eye-tracking feature.

6 Participating Teams & Systems

13 teams and a total of 42 participants registered on the competition website. All 13 teams, including 26 registered participants, submitted their predictions during the evaluation phase. Each team was allowed three submissions during the evaluation phase. Finally, 10 teams published system description papers outlining their approach (see Table 3 for all references).

Methods The participating teams submitted predictions generated from various approaches. Mainly two methods were used: (1) Boosting methods using tree-based algorithms with extensive feature extraction (e.g., CatBoost[2] or LightGBM[3]), and (2) neural network based approaches for regression such as fine-tuning transformer-based language models (Vaswani et al., 2017). Most teams achieved their best performance using an ensemble of predictors. Moreover, some teams also trained hybrid systems including both feature-based approaches and state-of-the-art language models.

Features The features included for training the systems include surface features (e.g., word length, sentence length, word positions in the sentence), lexical features (e.g., lemmas, named entities) token probability features (word frequency and n-gram metrics), syntactic features (e.g., part-of-speech tags and dependency parsing), text complexity metrics, behavioral measures, (e.g., concreteness, familiarity, age of acquisition), context features (i.e., information about the preceding and following tokens) as well as representations from state-of-the-art language models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019).

Additional data Only one team (Li and Rudzicz, 2021) used external eye-tracking data, leveraging the Provo corpus (Luke and Christianson, 2017) for additional word-level eye movement samples.

7 Results

In this section, we describe the prediction performance achieved by the participating teams. The official results of this shared task are presented in Table 3. The best results were achieved by a linguistic feature-based approach (Bestgen, 2021). As described above, other teams opted for neural

[2] https://catboost.ai/

[3] https://lightgbm.readthedocs.io/en/latest/

approaches (e.g., Li and Rudzicz, 2021 and Oh, 2021) or hybrid approaches (e.g., Yu et al., 2021 and Choudhary et al., 2021), combining linguistic features and state-of-the-art language representations.

The difficulty of predicting the individual eye-tracking features is analogous in all submitted systems. FFD is the most accurately predicted feature. This seems to suggest that the models are more capable to capture early processing stages of lexical access compared to late-stage semantic integration, indexed by TRT and NFIX.

Generally, the error for the three features representing reading times in milliseconds (FFD, GPT, and TRT), is much lower than for NFIX and FIXPROP. The latter are the features with the most variance. The mean baseline results also reveal the same patterns. The features with lower variance achieve lower MAEs. The FIXPROP feature, representing how likely a word is to be fixated, might be more challenging to predict since it is more dependent on subject-specific characteristics. Nevertheless, when comparing the MAEs of each eye-tracking feature to the mean baseline, the systems achieve the largest improvement on this feature.

8 Outlook & Conclusion

We presented the results of the first shared task on predicting token-level eye-tracking features recorded during natural sentences reading. We hope the CMCL Shared Task makes a lasting contribution to the field of linguistic cognitive modelling by providing researchers with a standard evaluation framework and a high quality dataset. Despite the limited size of the test set, many previously reached conclusions can now be tested more thoroughly and future models can be compared on a shared benchmark.

For future editions of this shared task, we see the following improvement opportunities: (1) providing an official development set during the training phase; (2) using additional metrics for assessment, such as R^2 to achieve a better understanding of the submitted models; (3) extending the dataset to include additional eye-tracking data from other English corpora, as well as including data from other languages such as Dutch or Russian (e.g., Cop et al., 2017 or Laurinavichyute et al., 2019).

References

Raksha Agarwal and Niladri Chatterjee. 2021. LangResearchLab_NC at CMCL2021 Shared Task: Predicting Gaze Behaviour using Linguistic Features and Tree Regressors. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Masayuki Asahara, Hajime Ono, and Edson T Miyamoto. 2016. Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING: Technical Papers*.

Christoph Aurnhammer and Stefan L Frank. 2019. Evaluating Information-theoretic Measures of Word Prediction in Naturalistic Sentence Reading. *Neuropsychologia*, 134:107198.

Alişan Balkoca, Abdullah Algan, Cengiz Acarturk, and Cagri Çöltekin. 2021. Team ReadMe at CMCL 2021 Shared Task: Predicting Human Reading Patterns by Traditional Oculomotor Control Models and Machine Learning. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data. In *Proceedings of ACL*.

Louise Gillian Bautista and Prospero Naval. 2020. Towards Learning to Read Like Humans. In *International Conference on Computational Collective Intelligence*, pages 779–791. Springer.

Yves Bestgen. 2021. LAST at CMCL 2021 Shared Task: Predicting Gaze Data During Reading with a Gradient Boosting Decision Tree Approach. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Shivani Choudhary, Kushagri Tandon, Raksha Agarwal, and Niladri Chatterjee. 2021. MTL782_IITD at CMCL 2021 Shared Task: Prediction of Eye-Tracking Features using BERT Embeddings and Linguistic Features. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting GECO: An Eyetracking Corpus of Monolingual and Bilingual Sentence Reading. *Behavior Research Methods*, 49(2):602–615.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text. In *Proceedings of NAACL*.

Franck Dary, Alexis Nasr, and Abdellah Fourtassi. 2021. TALEP at CMCL 2021 Shared Task: Non Linear Combination of Low and High-level Features

for Predicting Eye-Tracking Data. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Vera Demberg and Frank Keller. 2008. Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, 109(2):193–210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Susan E. Ehrlich and Keith Rayner. 1981. Contextual Effects on Word Perception and Eye Movements During Reading. *Journal of Verbal Learning and Verbal Behavior*, 20:641–65.

Victoria Fossum and Roger Levy. 2012. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the Human Sentence-processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

Adam Goodkind and Klinton Bicknell. 2018. Predictive Power of Word Surprisal for Reading Times Is a Linear Function of Language Model Quality. In *Proceedings of the LSA Workshop on Cognitive Modeling and Computational Linguistics*.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic Predictions of People Perusing: Evaluating Metrics of Language Model Performance for Psycholinguistic Modeling. In *Proceedings of the EMNLP Workshop on Cognitive Modeling and Computational Linguistics*.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing NLP with Cognitive Language Processing Signals. *arXiv preprint arXiv:1904.02682*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a Simultaneous EEG and Eye-tracking Resource for Natural Sentence Reading. *Scientific Data*.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In *Proceedings of LREC*.

Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. 2015. Integration and Prediction Difficulty in Hindi Sentence Comprehension: Evidence from an Eye-Tracking Corpus. *Journal of Eye Movement Research*, 8(2).

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Reinhold Kliegl, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the Mind During Reading: The Influence of Past, Present, and Future Words on Fixation Durations. *Journal of Experimental Psychology*, 135(1):12.

AK Laurinavichyute, Irina A Sekerina, SV Alexeeva, and KA Bagdasaryan. 2019. Russian Sentence Corpus: Benchmark Measures of Eye Movements in Reading in Cyrillic. *Behavior Research Methods*, 51(3):1161–1178.

Roger Levy. 2008. Expectation-based Syntactic Comprehension. *Cognition*, 106(3):1126–1177.

Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 Shared Task: RoBERTa with Multi-Stage Fine-Tuning for Eye-Tracking Prediction. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Steven G Luke and Kiel Christianson. 2017. The Provo Corpus: A Large Eye-tracking Corpus with Predictability Norms. *Behavior Research Methods*, pages 1–8.

Danny Merkx and Stefan L Frank. 2020. Comparing Transformers and RNNs on Predicting Human Sentence Processing Data. *arXiv preprint arXiv:2005.09471*.

Byung-Doh Oh. 2021. Team Ohio State at CMCL 2021 Shared Task: Fine-Tuned RoBERTa for Eye-Tracking Data Prediction. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Lavinia Salicchi and Alessandro Lenci. 2021. PIHKers at CMCL 2021 Shared Task: Cosine Similarity and Surprisal to Predict Human Reading Patterns. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. In *Proceedings of EMNLP*.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and Improving Natural Language Processing (in Machines) with Natural Language Processing (in the Brain). In *Advances in Neural Information Processing Systems*, pages 14928–14938.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Peter Vickers, Rosa Wainwright, Harish Tayyar Madabushi, and Aline Villavicencio. 2021. CogNLP-Sheffield at CMCL 2021 Shared Task: Blending Cognitively Inspired Features with Transformer-based Language Models for Predicting Eye Tracking Patterns. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv preprint arXiv:2006.01912*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

Qi Yu, Aikaterini-Lida Kalouli, and Diego Frassinelli. 2021. KonTra at CMCL 2021 Shared Task: Predicting Eye Movements by combining BERT with Surface, Linguistic and Behavioral Information. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.